Some Remarks on Sequence Design for DNA Computing (Abstract*)

Satoshi Kobayashi   Tomohiro Kondo
Department of Computer Science, University of Electro-Communications
1-5-1, Chofugaoka, Chofu, Tokyo 182-8585, JAPAN

1 Introduction

In this paper, we discuss on the design of DNA sequences for DNA computing([Adl94]). DNA strand design is one of the most important problems in order to obtain successful results of biological experiments. Furthermore, it is also important not only in DNA computing technology but in other biotechnologies such as the design of DNA chips.

Most of previous works introduced some variants of Hamming distance between sequences, and proposed methods to minimize the similarity between sequences based on that measure([GND97]). Typical and well known approaches contain combinatorial word design, random generation, and genetic algorithms, etc ([BKS00][DGR98][FLT97]). Recently, [AK02] proposed an interesting design method, called *template method*, for DNA word design.

This paper contain some results to further extend the template method. We propose to use multiple templates instead of single template in order to increase the number of DNA sequences to be designed. Further, we will report some theoretical properties of templates.

2 Template Method

In this paper, we focus mainly on the set of words over the alphabet either $\Sigma_{dna} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ or $\Sigma_{01} = \{0, 1\}$. The words over Σ_{01} are used by the template method ([AK02]) for designing DNA sequences over Σ_{dna}.

Let $x = x_1 \cdots x_n$ be a word on the alphabet Σ with $x_i \in \Sigma$ $(i = 1, ..., n)$. By $< x >$, we denote the proper subword of x of length $n-2$. In case of $n < 2$, $< x >$ is defined to be the empty word λ. The *reverse* of x, denoted by x^R, is

*The extended version of this paper will appear in Proc. of 8th International Meeing on DNA Based Computers, Sapporo, June, 2002.

the word $x = x_n \cdots x_1$. If $\Sigma = \Sigma_{dna}$, the *complement* of x, denoted by $\overline{x}$, is the word obtained by replacing each A in x by T and vice versa, and by replacing each C in x by G and vice versa. In case of $\Sigma = \Sigma_{01}$, $\overline{x}$ is the word obtained by replacing each 0 in x by 1 and vice versa.

Let $y = y_1 \cdots y_m$ be a word on the alphabet Σ with $y_i \in \Sigma$ $(i = 1, ..., m)$. In case of $n = m$, the Hamming distance $H(x, y)$ between x and y is the number of indices i such that $x_i \neq y_i$. For a finite set S of words of the same length, $H(S)$ is the minimum Hamming distance among all pairs of distinct elements in S.

In case of $n \leq m$, we define:

$$H_M(x, y) = min\{H(x, y') \mid y' \text{ is a subword of } y \text{ of length } n\}.$$

In case of $n > m$, $H_M(x, y)$ is defined to be n.

The template method is proposed to find a set of DNA sequences satisfying the following constraints. Let us consider the case when we are designing a set S of words over Σ_{dna} of the same length.

1. **Hamming distance** — For any pair of distinct words x, y in S, $H(x, y)$ should be at least a given integer d.

2. **Hamming distance with reverse-complement** — For any pair of (possibly the same) words x, y in S, $H(x, \overline{y}^R)$ should be at least a given integer d.

3. **Hamming distance between concatenated words** — For any (possibly the same) words x, y, z in S, $H_M(x, < yz >)$, $H_M(x, < \overline{y}^R \overline{z}^R >)$, $H_M(x, < y\overline{z}^R >)$, and $H_M(x, < \overline{y}^R z >)$ should be at least a given integer d.

4. **GC content** — The number of occurrences of G, C in a word x is called GC content of x. Then, this constraint requires every word in S should have the same GC content. GC content is an important indicator of the melting temperature of short oligonucleotides.

We define:

$$\begin{aligned} R(S) &= min\{\, H_M(x, \overline{y}^R), H_M(x, < yz >), H_M(x, < \overline{y}^R \overline{z}^R >), \\ &\qquad H_M(x, < y\overline{z}^R >), H_M(x, < \overline{y}^R z >) \mid x, y, z \in S\}, \\ ||S|| &= min\{H(S), R(S)\}. \end{aligned}$$

Then, the problem above can be formulated as follows:

Problem 1 For a given positive integer d and n, design a set S of n words over Σ_{dna} such that $||S|| \geq d$ and GC content of each word in S is the same to each

In order to solve this problem, the template method ([AK02]) uses a DNA sequence representation by a pair of binary words. Let us consider two homomorphisms $\phi, \psi : \Sigma_{dna}^* \rightarrow \Sigma_{01}$ such that $\phi(G) = \phi(C) = 1, \phi(A) = \phi(T) = 0$ and $\psi(A) = \psi(G) = 1, \psi(C) = \psi(T) = 0$. For a word w over Σ_{dna}, the binary represention of w is defined to be $(\phi(w), \psi(w))$, where $\phi(w)$ and $\psi(w)$ are called GC-component and AG-component of w, respectively [1]. Note that the binary representation of $\overline{w}^R$ is $(\phi(w)^R, \overline{\psi(w)}^R)$.

The idea of template method is summarized as follows. In the template method, GC-component of every word in S is fixed, and its unique GC-component is called GC-template of S [2]. The GC-template of S is carefully chosen so that $R(S) \geq d$ holds for a given d. On the other hand, AG-component of every word in S is designed so that $H(S) \geq d$ holds, for which we can use the theory of error correcting codes. Furthermore, since the GC-component of every word in S is fixed, the constraint of GC content is also satisfied.

In order to find a GC-template of S such that $R(S) \geq d$, it is useful to introduce the followig notation for a binary word x:

$$\begin{aligned} ||x|| &= min\{\, H(x, x^R), H_M(x, < xx >), H_M(x, < x^R x^R >), \\ &\quad H_M(x, < xx^R >), H_M(x, < x^R x >)\,\}. \end{aligned}$$

Then, for finding a GC-template of S such that $R(S) \geq d$, it suffices to find a binary word (GC-template) x such that $||x|| \geq d$. In [AK02], some theoretical analysis on the GC-templates was presented. Furthermore, for the length $l \leq 30$, all of the best GC-templates are searched exhaustively, and presented at the appendix.

3 Using Multiple Templates

In this section, we will discuss on the effectiveness of using multiple templates. Let T be a finite set of binary words of length n to be used as templates. Then, we define:

$$\begin{aligned} R'(T) &= min\{\, H(x, y^R), H_M(x, < yz >), H_M(x, < y^R z^R >), \\ &\quad H_M(x, < yz^R >), H_M(x, < y^R z >) \mid x, y, z \in T\}, \\ ||T|| &= min\{H(T), R'(T)\}. \end{aligned}$$

It is straightforward to see the following facts:

Fact 1 For any set T of binary words such that $||T|| \geq d$ and any set S of words over Σ_{dna} using T as a set of GC-templates, $R(S) \geq d$ holds.

[1] In [AK02], the bitwise negation of $\phi(w)$ is called a template of w, and $\psi(w)$ is called a code of w.

[2] In [AK02], the bitwise negation of GC-template is just called as template.

In Table 1, we summarize the obtained results on the maximum $||T||$ values of multiple templates, where the number of words of $T$ is 2. The integers in the round brackets are the number of $T$ that gives the maximum $||T||$ values. In the representation $[x, y]$, $x$ and $y$ are lower and upper bounds, respectively. The suffix $g$ represents that this lower bound was obtained by applying genetic algorithms.

length	3	4	5	6	7	8	9	10	11	12
$\|\|T\|\|$	– –	– –	– –	[0,1]	1	2 (8)	2 (496)	2	[2,3]	4 (8)
length	13	14	15	16	17	18	19	20	21	22
$\|\|T\|\|$	4 (88)	4 (354)	4	[4,5]	5	6 (60)	6^g	[6^g,7]	[6,7]	[7,8]
length	23	24	25	26	27	28	29	30		
$\|\|T\|\|$	[7^g,8]	8	8	9	9	[8,10]	[9^g,10]	[10,11]		

Table 1: Maximum $||T||$ values of multiple Templates $T$ of size 2

In case of $l = 18$, we can, for instance, choose as a set of AG-components, a code of length 18, Hamming distance at least 6, and constant weight 9, which consists of 304 code words(Table I-B in [BSS90]). Therefore, we can design 608 DNA sequences. In case of $l = 24$, we can, for instance, choose as a set of AG-components, a code of length 24, Hamming distance at least 8, and constant weight 12, which consists of 2576 code words(Table I-C in [BSS90]), which produces 5152 DNA sequences. Although we should further check by biological experiments whether $\frac{l}{3}$ mismathes are enough or not, these examples demonstrate the potential effectiveness of multiple templates method for not all but some selected lengths.

4 Some Properties of Templates

In this section, we give a theoretical analysis on the properties of GC-templates.

Let us consider the behavior of the following function:

$$\mu(l) \quad =_{def} \quad max\{\frac{||x||}{l} \mid x \in \Sigma^l\}.$$

By the same discussion as in Lemma 2.5 of [AK02], we can derive the following theorem.

Theorem 1 For any $l$, $\mu(l) \leq \frac{1}{2}$ holds.

Furthermore, we can show the following main theorem.

Theorem 2 For infinitely many $l$'s, $\mu(l) \geq \frac{11}{30}$ holds.

5 Conclusions

This paper presented some results to further extend the template method. We proposed to use multiple templates method, which was shown to have potential abiity to increase the number of DNA sequences to be designed. Finally, we derived some theoretical properties of templates.

However, we have the following problems to be studied in the future works. At first, the validity of the proposed template method should be checked by biological experiments. In particular, it should be checked whether $\frac{1}{3}$ mismatches of length l sequences is enough or not. Although some properties of templates are presented in the current paper, most of the important properties have not been revealed yet. In particular, the reason why the best templates of length l might have approximately $\frac{l}{3}$ mismatches is not clear. Finally, the proposed method does not guarantee the optimality in the number of sequences to be designed, or in the number of mismaches between sequences. Further theoretical analysis should be done in a more general framework.

Acknowledgement

The first author was supported in part by Japan Society for the Promotion of Science, Grant-in-Aid for Exploratory Research 13878054, 2001.

References

[Adl94] L. Adleman, Molecular Computation of Solutions to Combinatorial Problems. *Science* **266**, pp.1021-1024, 1994.

[AK02] M. Arita and S. Kobayashi, DNA Sequence Design Using Templates, *New Generation Computing*, to appear.

[BKS00] A. Ben-Dor, R. Karp, B. Schwikowski, and Z. Yakhini, Universal DNA Tag Systems: A Combinatorial Design Scheme, Proc. of the 4th Annual International Conference on Computational Molecular Biology (RECOMB2000), pp.65-75, 2000.

[BSS90] A.E. Brouwer, J.B. Shearer, N.J.A. Sloane, and W.D. Smith, A New Table of Constant Weight Codes, *IEEE Trans. on Information Theory*, **36**, pp.1334-1380, 1990.

[DGR98] R. Deaton, M. Garzon, J.A. Roze, D.R. Franceschetti, R.C. Murphy and S.E. Jr. Stevens, Reliability and Efficiency of a DNA Based Computation, *Physical Reviw Letter*, **80**, pp.417-420, 1998.

[FLT97] A. G. Frutos, Q. Liu, A. J. Thiel, A. M. W. Sanner, A. E. Condon, L. M. Smith and R. M. Corn, Demonstration of A Word Design Stragety for DNA Computing on Surfaces, Nucleic Acids Research, Vol.25, No.23, pp.4748-4757, 1997.

[GND97] M. Garzon, P. Neathery, P. Deaton, R.C. Murphy, D.R. Franceschetti, and S.E. Jr. Stevens, A New Metric for DNA Computing, In *Proc. of 2nd Annual Genetic Programming Conference*, Morgan Kaufmann, pp.472-478, 1997.

[MS77] E. J. MacWilliams and N.J.A. Sloane, The Theory of Error-Correcting Codes, North-Holland, 1977.